Analyses of protein cores reveal fundamental differences between solution and crystal structures

RESEARCH ARTICLE Proteins: Structure, Function, and Bioinformatics
DOI10.1002/prot.25884

Title: Analyses of protein cores reveal fundamental differences between solution and crystal structures

Authors:

Zhe Mei[1,2], John D. Treado[1,3], Alex T. Grigas[1,4], Zachary A. Levine[5,6], Lynne Regan[7], Corey S. O'Hern[1,3,8,9]

Affiliations:

[1]Integrated Graduate Program in Physical & Engineering Biology, Yale University, New Haven, Connecticut 06520, USA

[2]Department of Chemistry, Yale University, New Haven, Connecticut 06520, USA

[3]Department of Mechanical Engineering & Materials Science, Yale University, New Haven, Connecticut 06520, USA

[4]Graduate Program in Computational Biology & Bioinformatics, Yale University, New Haven, Connecticut 06520 USA

[5]Department of Pathology, Yale University, New Haven, Connecticut 06520, USA

[6]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, 06520

[7]Institute of Quantitative Biology, Biochemistry and Biotechnology, Center for Synthetic and Systems Biology, School of Biological Sciences, University of Edinburgh

[8]Department of Physics, Yale University, New Haven, Connecticut 06520, USA

[9]Department of Applied Physics, Yale University, New Haven, Connecticut 06520, USA

Abstract

Received: Nov 27, 2019;Revised: Feb 05, 2020;Accepted: Feb 23, 2020

There have been several studies suggesting that protein structures solved by NMR spectroscopy and x-ray crystallography show significant differences. To understand the origin of these differences, we assembled a database of high-quality protein structures solved by both methods. We also find significant differences between NMR and crystal structures—in the root-mean-square deviations of the C_α atomic positions, identities of core amino acids, backbone and side chain dihedral angles, and packing fraction of core residues. In contrast to prior studies, we identify the physical basis for these differences by modelling protein cores as jammed packings of amino-acid-shaped particles. We find that we can tune the jammed packing fraction by varying the degree of thermalization used to generate the packings. For an athermal protocol, we find that the average jammed packing fraction is identical to that observed in the cores of protein structures solved by x-ray crystallography. In contrast, highly thermalized packing-generation protocols yield jammed packing fractions that are even higher than those observed in NMR structures. These results indicate that thermalized systems can pack more densely than athermal systems, which suggests a physical basis for the structural differences between protein structures solved by NMR and x-ray crystallography.

Section 1. Introduction

It is generally acknowledged that protein structures determined by x-ray crystallography versus NMR exhibit small but significant differences. It is by no means resolved, however, whether these differences stem from differences in the experimental methods themselves, or if they reflect physical differences in proteins under the different conditions in which the measurements are made [1, 2, 3, 4, 5, 6, 7, 8]. To begin to answer this question, one must directly compare high-quality structures of the same protein solved by both methods. Choosing x-ray crystal structures based on their resolution is a straightforward way to identify well-specified structures. In our

database of structures solved by both x-ray crystallography and NMR, we only include structures that have been solved by x-ray crystallography at a resolution of 2.1Å or less. We also show that our results do not depend on this resolution threshold as long as it is 3Å or less.

There is, however, no universally accepted metric to assess the quality of NMR structures. We therefore defined one; we determined the number of NMR restraints per residue beyond which structures do not change significantly with the addition of more restraints, and only used structures with at least this number of restraints per residue on average. (See Fig. 1.) Applying these selection criteria, we created a data set of 21 proteins whose structures have been determined by both x-ray crystallography and NMR. We created an additional dataset of 51 high-quality NMR protein structures (defined in the same way), for which there is no companion x-ray crystal structure, in an attempt to exclude any influence of 'crystallizability' on the NMR protein structures. In addition, as a reference set of high-resolution protein structures solved by x-ray crystallography, we use a dataset of 221 high-resolution protein structures collected by Wang and Dunbrack [9]. Finally, we created a dataset of structures that have been solved multiple times by x-ray crystallography, with resolution of 2.0Å or less and the same crystal forms and space groups, to allow us to assess structural variations that arise from thermal fluctuations.

We find that the root-mean-square deviations (RMSD) of the positions of core C_α atoms within an NMR 'bundle' is greater than the RMSD of core C_α atoms of the set of protein crystal structures that have been solved multiple times, a result found by researchers in prior work [1]. Also, the difference between an x-ray crystal structure and each structure in the NMR 'bundle' is greater than the spread within the NMR bundle. To gain deeper insight into these differences, we performed side chain repacking studies on core residues in both x-ray crystal and NMR structures using the hard-sphere plus stereochemical constraint model developed in our previous work [10, 11]. We find that the hard-sphere plus stereochemical constraint model can predict the side chain dihedral angle conformations of core residues equally well in both NMR and x-ray crystal structures, predicting $\Delta\chi$ values to within 30° of the experimental structures. In our

previous work, we found that the predictability of side chain conformations is strongly correlated with the local packing fraction ϕ, i.e. where we obtain almost 100% prediction accuracy of side chain conformations for core residues with packing fraction $\phi \geq 0.55$. We therefore also calculate the core packing fractions in NMR and x-ray crystal structures, and find that the cores of NMR structures are more tightly packed than the cores of x-ray crystal structures [1].

To further explore the physical basis for these observations, we generated jammed packings of amino-acid-shaped particles computationally, and determined whether we can tune their packing fraction using protocols with different degrees of thermalization. We find that depending on the thermalization protocol we use, we can match the packing fraction to that which we observe in the cores of structures determined by x-ray crystallography and NMR. Specifically, the packing fraction of amino acid-shaped particles in the athermal limit corresponds to that in the cores of protein crystal structures, whereas the packing fraction we observe in cores of NMR structures is higher, but less than that achieved in the limit of strong thermalization. Thus, the core packing fraction for protein structures determined by x-ray crystallography and NMR are both physically reasonable, and we speculate that the higher packing fraction for NMR structures reflects the different conditions under which NMR structures are determined.

Section 2. Methods

Section 2.1 Protein structure datasets

All experimental proteins used in this study were culled from the RCSB Protein Data Bank (PDB). We used datasets of (a) high resolution crystal structures, (b) x-ray crystal-NMR structure pairs, (c) duplicate x - ray crystal structures, (d) high-quality, non-paired NMR structures, (e) mutated crystal structures, and (f) structural prediction decoys from the 12th Critical Assessment of Protein Structure Prediction (CASP12). We show the full PDB id's in the Supplementary Information (SI) for all datasets except the high-resolution crystal structures and the CASP12

decoys and targets. Detailed descriptions of the datasets are provided in the Supplementary Information.

Section 2.2. NMR structural quality

There is no universally accepted metric to assess the quality of NMR structures [2]. To define one, we determined the number of NMR restraints per residue beyond which the structures do not change significantly with the addition of more restraints. We measured the root-mean-square deviation (RMSD) of the C_α positions of a given set of residues defined by their sequence location on two models *i* and *j* within an NMR bundle:

$$\Delta(i,j) = \sqrt{\frac{1}{N_S}\sum_{\mu=1}^{N_S}\left(\vec{c}_{\mu,j} - \vec{c}_{\mu,i}\right)^2}\ , (1)$$

where $\vec{c}_{\mu,i}$ is the position of the C_α atom on residue μ in model *i*, and N_S is the number of residues being compared. We can calculate the average RMSD $\langle\Delta(i\ ,j\ )\rangle$ by averaging over all pairs of models i and j. As shown in Fig. 1, Δ plateaus to a value near 1.5 Å when the average number of restraints per residue reaches $N_r \geq 15$. Thus, we restrict our NMR datasets (Tables S1 and S3 in the SI) to proteins for which the NMR structures possess on average ≥ 15 restraints per residue.

Section 2.3 Relative solvent accessible surface area (rSASA)

We define core residues based on their solvent-accessible surface area (SASA). To calculate the SASA, we use the NACCESS software package [12] that implements an algorithm originally proposed by Lee and Richards [13]. The algorithm takes *z* -slices of the protein, determines the solvent accessibility of the sets of contours using a probe molecule of a given radius, and integrates the SASA over the slices. We use a water-molecule-sized probe with radius 1.4 Å and z-slices with thickness $\Delta z = 10^{-3}$ Å, which were used in previous work [11]. We calculate the SASA for a given residue μ in both the context of the surrounding protein ($SASA_\mu^{context}$) and for

the residue “extracted" from the protein and modeled as a dipeptide mimetic ($SASA_{\mu}^{dipeptide}$), with all bond lengths, bond angles, and dihedral angles preserved. We define the relative SASA ($rSASA_{\mu}$) for residue μ as the ratio

$$rSASA_{\mu} = \frac{SASA_{\mu}^{context}}{SASA_{\mu}^{dipeptide}}. (2)$$

We define core residues as those with $rSASA < 10^{-3}$, which was found in previous work [11] to be the largest value of $rSASA$ such that the packing fraction and side chain repacking predictability no longer depend on the value of the $rSASA$ cutoff when it is decreased.

Section 2.4 Packing fraction

The most direct way to characterize packing in protein cores is to measure the dimensionless volume fraction, or packing fraction ϕ. The packing fraction ϕ_{μ} of a single residue μ in a protein core is defined as

$$\phi_{\mu} = \frac{v_{\mu}}{V_{\mu}^{v}}, (3)$$

where v_{μ} is the volume of residue μ, and V_{μ}^{v} is the volume of the Voronoi cell surrounding residue μ. To calculate the Voronoi tessellation for a given protein core, we employ surface Voronoi tessellation [14], which defines a Voronoi cell as the region of space in a given system that is closer to the bounding *surface* of residue μ than to the bounding surface of any other residue in the system. We calculate the surface Voronoi tessellations using the POMELO software package [15]. This software approximates the bounding surfaces of each residue by triangulating points on the residue surfaces. We find that using ~ 400 points per atom, or ~ 6400 surface points per residue, gives an accurate representation of the surface Voronoi cells and the results do not change if more surface points are included. Note that to calculate the average packing fraction of a protein core, we define

$$\langle \phi \rangle = \frac{\sum_\mu v_\mu}{\sum_\mu V_\mu^v}, (4)$$

where the sum over μ includes only core residues. In this work, we define a protein core as the set of residues with $rSASA < 10^{-3}$ that all share at least one surface Voronoi cell face with each other.

Section 2.5 Side chain repacking

To better understand the dominant forces determining the side chain conformations in protein cores, we have developed a protocol that can repack the side chains of core residues assuming that the non-bonded atomic interactions are hard- sphere-like, and that bond lengths and angles are tightly constrained around experimentally-observed values. The hard-sphere plus stereochemical constraint model has been used extensively in previous work (e.g. Refs. [10, 11] and references therein) to accurately place hydrophobic residue side chains in the cores of the crystal structures of globular proteins, transmembrane proteins, and protein-protein interfaces. In this model, we sample all possible combinations of the side chain dihedral angles of the core residues, and calculate the purely repulsive Lennard-Jones interaction energy (Eq. (6)) between non-bonded atoms for each combination. The backbone dihedral angles of each core residue are fixed to their experimental values, as well as the side chain and backbone dihedral angles of the rest of the protein. We obtain a probability distribution for the side chain dihedral angle combinations of each core residue using Boltzmann weighting, and repeat this procedure over an ensemble of structures with core residues given different bond-length and bond-angle variants constrained around the experimental values. We then average the probability distributions over this ensemble and identify the side chain dihedral angle combination with the highest probability. We employ this model to study residue packing and side chain placement in the cores of both x-ray and NMR structures. Additional details of the method are given in the SI.

Section 2.6 Jammed packings of amino-acid-shaped particles

In previous work [16], we found that the packing fraction and void distribution of protein cores is well-modeled by computer simulations of jammed packings of purely repulsive, rigid, and non-backbone-connected particles shaped like hydrophobic residues. The amino-acid-shaped particles include the backbone N, C_α, C, and O atoms, as well as all side chain atoms and hydrogens placed using the REDUCE software [17]. Atomic radii are listed in Table S6 in the SI. To prepare the jammed packings, we first place N amino-acid-shaped particles with random positions and orientations in a cubic box with periodic boundary conditions at an initially dilute packing fraction $\phi_0 = 0.1$. The packing fraction is increased by small steps $\Delta\phi$, with each followed by energy minimization, to mimic athermal isotropic compression of the system. We also carry out thermalized compression protocols, where we thermalize the amino-acid shaped particles between compression steps. In this method, we run molecular dynamics trajectories at constant temperature T for a fixed amount of time t_{MD}, and then minimize the total potential energy of the system U using the FIRE minimization method [18] prior to the next compression step. We terminate the packing generation protocols when the minimized potential energy per particle satisfies $10^{-16} < U/N\epsilon \leq 2 \times 10^{-16}$, where ϵ is the energy scale of the non-bonded atomic interactions, and the kinetic energy per particle $K/N\epsilon < 10^{-30}$. Further details of the packing-generation protocols are given in the SI.

Section 3. Results

We first compare pairs of structures determined by x-ray crystallography and NMR spectroscopy by quantifying the root-mean-square deviation (RMSD, Eq. (1)) of the C_α positions of a given set of residues defined by their sequence location on two structures i and j. For the NMR datasets, i and j represent each model within a bundle, and, for the x-ray crystal duplicate dataset, i and j represent each of the duplicates. As mentioned in Sec 2.3, we define core residues as residues

with small ($< 10^{-3}$) relative solvent-accessible surface area (rSASA), as defined in Eq. (2) in Sec. 2.3. In Fig. 2 (a), we compare the distributions $P(\Delta_{core})$ of RMSD values of core residues in x-ray crystal structure duplicates and RMSD values of core residues in NMR bundles. We show that the fluctuations among x-ray crystal structure duplicates are consistent with B-factor fluctuations of the C_α positions of core residues, B, which are given by $\Delta = \sqrt{3B/8\pi^2}$ .We also compare x-ray crystal and NMR structures for the same proteins by calculating the RMSD between C_α atoms of core residues.

To quantify differences between each RMSD distribution, we compute the Jensen-Shannon (JS) divergence [19] for each distribution in Fig. 2 (a). The JS divergence between the x-ray duplicate RMSD distribution and the B-factor distribution is 0.5, while the JS divergence between the NMR intrabundle RMSD and the NMR-x-ray RMSD is 1.1, which demonstrates that the RMSD between NMR and x-ray structures is greater than the RMSD differences within a bundle of NMR structures, or between duplicate x-ray structures of the same protein. Because x-ray duplicate RMSD values are similar to B-factor RMSD values, the relatively low JS divergence indicates that fluctuations across duplicate crystal structures is dominated by the uncertainty in atomic positions arising from thermal motion. Whereas the larger JS divergence between NMR intrabundle RMSD and NMR-x-ray RMSD values, as well as the broad tail in the NMR-x-ray RMSD distribution, suggests that differences between structures solved by both NMR and crystallography are larger than those expected in both the ensemble of x-ray structures and in NMR bundles individually. That is, while the fluctuations in the ensemble of observed NMR structures is larger than those in the observed ensemble of crystal structures, these two ensembles typically occupy distinct, non-overlapping regions of configuration space.

We also calculate the side chain dihedral angle fluctuations $\Delta\chi$ for the same pairs of structures; we define $\Delta\chi(\mu|i,j)$ as the distance between the side chain conformations of residue μ within structures i and j, i.e.

$$\Delta\chi(\mu|i,j) = \sqrt{\left(\vec{\chi}_{\mu,j} - \vec{\chi}_{\mu,i}\right)^2}\,, (5)$$

where $\vec{\chi}_{\mu,i}$ is the set of side chain dihedral angles $(\chi_1, \ldots, \chi_m)$ for residue μ on structure i. Note that in Fig. 2 (b), we measure $\Delta\chi$ between two experimental structures of the same protein, whereas in Fig. 3 (a) and (b) we measure $\Delta\chi$ between an experimental structure and a prediction using the hard-sphere plus stereochemical constraint model.

In Fig. 2, we show that the conformations of both the backbone and side chains of core residues fluctuate less in x-ray crystal structures compared to the conformations within an NMR bundle, but that the fluctuations within an NMR bundle are smaller than those *between* the x-ray crystal and NMR structure pairs [1, 7, 8]. The inset to Fig. 2 (b) illustrates the proportion of configuration space sampled for structures solved by both NMR and x-ray crystallography. Structures determined by x-ray crystallography sample states in a relatively small volume of configuration space compared to that sampled by structures in an NMR bundle. Moreover, these two ensembles are separated by a characteristic distance that is larger than the scale of fluctuations in either ensemble.

To put these structural differences in context, we also analyze fluctuations in a database of pairs of x-ray crystal structures of wild-type proteins and the same protein with a single core mutation and also high-scoring submissions from a recent Critical Assessment of Protein Structure Prediction (CASP) competition [20]. In the SI (see Fig. S3), we show that the fluctuations of single-site core mutants relative to wildtype structures is similar to that in x-ray crystal structure duplicates. In contrast, submissions to CASP12 exhibit much larger fluctuations. Because CASP12 submissions are computational predictions, not experimentally determined structures, one might expect larger fluctuations. The fluctuations among CASP12 submissions is also larger than those between structures of the same protein determined by x-ray crystallography or NMR. In the SI, we report additional measures of structural fluctuations, such as fluctuations in identities of core residues (Fig. S2). We also show in Figs. S4 and S5 that the

global and core RMSD of the C_α positions do not depend on the resolution of the x-ray crystal structures, as long as the resolution is less than 3Å.

To understand the origin of differences between x-ray crystal and NMR structures, we investigated if these differences are due to physical forces governing sidechain placement of core residues. In previous work, we showed that the hard-sphere plus stereochemical constraint model uniquely specifies the sidechain dihedral angles of core residues in protein crystal structures [11]. One potential source of differences in fluctuations in NMR and crystal structure cores could be that the cores in NMR structures are less well-resolved, and the sidechains are poorly placed due to insufficient information to uniquely define their conformations. Such methodological inaccuracies have been suggested by previous studies, where computational refinement moves NMR backbone and sidechain dihedral angles towards those of x-ray crystal structures [1, 2, 3, 4]. However, as shown in Fig. 3 (a) and (b), we find that we can repack sidechains of core residues in NMR structures just as accurately as we can repack the same sidechains in high-resolution x-ray crystal structures. The side chain repacking protocol is described in Sec. 2.5 and in further detail in the SI. For side chain repacking, we calculate the repulsive Lennard-Jones potential energy of overlap U_{RLJ} between side chains of core residues in the pairs of structures. The potential energy of a single residue μ with side chain confirmation $\vec{\chi}_\mu$ is defined by

$$U_{RLJ} = \sum_{\nu}^{N} \sum_{i,j} \frac{\epsilon}{72} \left[1 - \left(\frac{\sigma_{ij}^{\mu\nu}}{r_{ij}^{\mu\nu}} \right)^6 \right]^2 \Theta\left(\sigma_{ij}^{\mu\nu} - r_{ij}^{\mu\nu} \right) , (6)$$

where the potential energy is evaluated as a sum over all non-bonded atomic interactions. $r_{ij}^{\mu\nu}$ is the distance between atoms i and j on residues μ and ν, $\sigma_{ij}^{\mu\nu} = (\sigma_i^\mu + \sigma_j^\nu)/2$, and σ_j^μ is the diameter of atom i on residue μ. The Heaviside step function Θ enforces the potential to be purely-repulsive. We find that the distribution of repulsive Lennard-Jones energies between core side chains are almost identical when comparing x-ray crystal and NMR structures, which

indicates that the NMR and crystal structure cores are statistically at the same energies. (See Fig. 3 (c).)

However, when we investigate the packing fraction ϕ of core residues for x-ray crystal and NMR structures, we find important differences. In Fig. 4, we plot the probability distribution $P(\varphi)$ of the packing fraction for core residues in x-ray crystal and NMR structures. The average packing fraction of core residues in the protein structures in the datasets determined by x-ray crystallography is $\langle\phi\rangle = 0.55 \pm 0.01$, a value that is consistent with our previous results for the packing fraction of core residues in globular and transmembrane protein cores and the cores of protein-protein interfaces solved by x-ray crystallography [11, 16]. For core residues of protein structures in the NMR database, the average packing fraction is higher with $\langle\phi\rangle = 0.59 \pm 0.02$. We believe that this is the first time that such a difference in the packing fraction of core residues in high-quality protein structures determined by both x-ray crystallography and NMR has been reported.

We were concerned that the higher packing fraction of core residues in protein structures determined by NMR could be an artifact of improperly-placed sidechains that overlap with neighboring residues, which would artificially increase the observed packing fraction. However, comparison of the distribution of overlap energies measured by U_{RLJ} (Eq. (6)) in Fig. 3 (c) demonstrates that the two methods result in almost identical energies, and therefore almost identical atomic overlaps. The difference in the packing fraction of core residues was at first surprising, because our previous studies showed that the cores of x-ray crystal structures pack as densely as jammed packings of purely-repulsive amino-acid-shaped particles without backbone constraints generated using a protocol of successive compressions followed by energy minimization [21, 16].

We therefore revisited the protocol with which we prepared jammed packings of amino-acid-shaped particles [16]. In our previous work, packings were prepared using an “athermal” protocol, where kinetic energy was drained rapidly from the system during the packing

preparation. For the athermal protocol, amino acids were initialized in a cubic simulation box at a small initial packing fraction ϕ_0 and compressed by small increments $\Delta\phi$ with each followed by energy minimization (see Sec. 2.6 and SI for additional details.) Because the amino-acid-shaped particles were not allowed to translate and rotate significantly between each compression step, the jammed packings at $\phi \approx 0.55$ were obtained at the first metastable jammed state that the protocol encounters. However, the packing fractions that can be achieved in packings of amino-acid-shaped particles are protocol-dependent; we next investigated more thermalized protocols to see how different protocols lead to different jammed packing fractions.

We chose a family of annealing packing-generation protocols. We initialize the system in a dilute configuration, and compress the system in small increments $\Delta\varphi$ between periods of molecular dynamics simulations of purely repulsive amino acids-shaped particles in the canonical ensemble for a time period t_{MD} at thermal energy k_BT. (See SI for details.) We find that temperature only acts to renormalize t_{MD}, i.e. a longer simulation at a lower temperature will yield the same results as a shorter simulation at higher temperatures. Thus, there is another time scale associated with the annealing protocol, $t_{QA} = c(T)t^*$, where $c(T)$ is a dimensionless quantity that depends on temperature, $t^* = \sqrt{\frac{m_R\sigma_R^2}{\epsilon}}$ and m_R and σ_R are the mass and diameter of the smallest residue. We find that plotting the ensemble-averaged packing fraction $\langle\phi\rangle$ of jammed packings of amino acid-shaped particles versus $\tau = \frac{t_{MD}}{t_{QA}} = n\left(\frac{k_BT}{\epsilon}\right)^{\alpha}$, collapses the data for different temperatures and time periods onto a single curve (Fig. 5). The exponent $\alpha = 0.4 \pm 0.01$ and n is the number of time steps between compression increments.

Two limits of packing fractions emerge over the range of annealing protocols we tested; an athermal limit, which corresponds to packing fractions in cores of x-ray crystal structures [11], and the thermalized limit with $\langle\phi\rangle \approx 0.62$. The packing fraction in the cores of protein structures solved by NMR fall between these two extremes with $\langle\phi\rangle = 0.59$. The states at exceedingly high

packing fractions exist only in the limit of extremely long annealing times. The results of simulations using different protocols are consistent with the differences observed in cores of protein structures solved by x-ray crystallography and NMR. The fact that thermalized packing protocols yield NMR-like packing fractions, and that athermal protocols generate x-ray crystal-like packing fractions, suggests that fluctuations are distinct for these two methods.

Section 4. Discussion & Conclusions

In this work, we compare the fluctuations of protein structures characterized by both NMR and x-ray crystallography, and find several key results: first, we found that RMSD values between core residues in duplicated x-ray crystal structures are smaller than RMSD values between core residues across multiple structures in NMR bundles, but these RMSD values are still smaller than the RMSD values between core residues in NMR and x-ray crystal structure pairs. These findings suggest that NMR and x-ray crystal structures occupy distinct regions in configuration space. However, we also showed that the hard-sphere plus stereochemical constraint model is extremely accurate in side chain conformation prediction for core residues in both x-ray crystal and NMR protein structures. Measurements of the core packing fraction show that NMR structures possess denser cores, even though the cores in x-ray crystal and NMR structures possess the same overlap energy. To resolve this apparent discrepancy, we prepare jammed packings of amino-acid-shaped particles both athermally and with thermal agitation, and find that packings produced in the athermal limit resemble the cores of x-ray crystal structures, while thermalized packings resemble cores in NMR structures. This result suggests that there are subtle yet real differences in the fluctuations between structures characterized by x-ray crystallography and NMR spectroscopy. The fluctuations are larger in NMR structures than in x-ray crystal structures, and these fluctuations lead to slightly denser packing in the core.

A previous study that also compared protein structures determined by x-ray crystallography and NMR suggested that the crystal environment restricts dynamical fluctuations, whereas bundles of NMR structures in solution contain the full dynamics one would expect from elastic network models for proteins [6]. The work we present here provides further evidence to support this conclusion, but whether the differences are due to crystalline contacts [6, 7, 22] or the different temperatures at which the protein structures are characterized [23] remains to be determined. Interestingly, several structures used in our dataset of duplicate crystal structures were resolved at room temperature (~ 300 K), as opposed to the cryogenic temperatures typically used in x-ray crystallography. We found that core RMSD values do not change significantly when considering duplicate x-ray crystal pairs solved at different temperatures, which suggests that the crystal environment is the dominant cause of the differences between structures solved by NMR and x-ray crystallography. To fully resolve this question, however, further characterization of protein structure fluctuations at different temperatures is required.

Acknowledgements

The authors acknowledge support from NIH training Grant No. T32EB019941 (J.D.T.), the Integrated Graduate Program in Physical and Engineering Biology (Z.M.), and NSF Grant No. PHY-1522467 (C.S.O.). This work also benefited from the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center. We thank Pat Loria and Peter Moore for helpful discussions.

References

[1] Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein AV, Galzitskaya OV. Comparison of X-ray and NMR structures: Is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? Proteins: Structure, Function, and Bioinformatics 2005;60(1):139–147.

[2] Schneider M, Fu X, Keating AE. X-ray vs. NMR structures as templates for computational protein design. Proteins: Struc- ture, Function, and Bioinformatics 2009;77(1):97–110.

[3] Mao B, Tejero R, Baker D, Montelione GT. Protein NMR structures refined with Rosetta vave higher accuracy relative to corresponding X-ray crystal structures. Journal of the American Chemical Society 2014 02;136(5):1893–1906.

[4] Koehler Leman J, D'Avino AR, Bhatnagar Y, Gray JJ. Comparison of NMR and crystal structures of membrane pro- teins and computational refinement to improve model quality. Proteins: Structure, Function, and Bioinformatics 2018;86(1):57–74.

[5] Best RB, Lindorff-Larsen K, DePristo MA, Vendruscolo M. Relation between native ensembles and experimental struc- tures of proteins. Proceedings of the National Academy of Sciences 2006;103(29):10901–10906.

[6] Yang LW, Eyal E, Chennubhotla C, Jee J, Gronenborn AM, Bahar I. Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. Structure 2007;15(6):741 – 749.

[7] Sikic K, Tomic S, Carugo O. Systematic comparison of crystal and NMR protein structures deposited in the Protein Data Bank. The Open Biochemistry Journal 2010;4(83-95):83–95.

[8] Everett JK, Tejero R, Murthy SBK, Acton TB, Aramini JM, Baran MC, et al. A community resource of experimental data for NMR / X-ray crystal structure pairs. Protein Science 2016;25(1):30–45.

[9] Wang G, Dunbrack RL Jr. PISCES: A protein sequence culling server. Bioinformatics

2003;19(12):1589–1591.

[10] Caballero D, Virrueta A, O'Hern CS, Regan L. Steric interactions determine side-chain conformations in protein cores. Protein Engineering, Design and Selection 2016;29(9):367–376.

[11] Gaines JC, Acebes S, Virrueta A, Butler M, Regan L, O'Hern CS. Comparing side chain packing in soluble proteins, protein- protein interfaces and transmembrane proteins. Proteins: Structure, Function, and Bioinformatics 2018;86(5):581–591.

[12] Hubbard SJ, Thornton JM, NACCESS; 1993. http://wolf.bms.umist.ac.uk/naccess/.

[13] Lee B, Richards FM. The interpretation of protein structures: Estimation of static accessibility. Journal of Molecular Biology 1971;55(3):379–400.

[14] M Schaller F, Kapfer S, Evans M, J F Hoffmann M, Aste T, Saadatfar M, et al. Set Voronoi diagrams of 3D assemblies of aspherical particles. Philosophical Magazine 2013;93:3993–4017.

[15] Weis S, Schönhöfer PWA, Schaller FM, Schröter M, Schröder-Turk GE. Pomelo, A tool for computing Generic Set Voronoi Diagrams of Aspherical Particles of Arbitrary Shape. EPJ Web Conf 2017;140:06007.

[16] Treado JD, Mei Z, Regan L, O'Hern CS. Void distributions reveal structural link between jammed packings and protein cores. Phys Rev E 2019;99:022416.

[17] Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. Journal of Molecular Biology 1999;285(4):1735 – 1747.

[18] Bitzek E, Koskinen P, Gähler F, Moseler M, Gumbsch P. Structural Relaxation Made Simple. Phys Rev Lett 2006;97:170201.

[19] Endres DM, Schindelin JE. A new metric for probability distributions. IEEE Transactions on Information Theory 2003 July;49(7):1858–1860.

[20] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of

methods of protein structure prediction (CASP)—Round XII. Proteins: Structure, Function, and Bioinformatics 2018 2;86(S1):7–15.

[21] Gaines JC, Smith WW, Regan L, O'Hern CS. Random close packing in protein cores. Phys Rev E 2016;93:032415.

[22] Halle B. Biomolecular cryocrystallography: Structural changes during flash-cooling. Proceedings of the National Academy of Sciences 2004;101(14):4793–4798.

[23] Fraser JS, van den Bedem H, Samelson AJ, Lang PT, Holton JM, Echols N, et al. Accessing protein conformational ensembles using room-temperature X-ray crystallography. Proceedings of the National Academy of Sciences 2011;108(39):16247–16252.

[24] Wang G, Dunbrack RL Jr. PISCES: Recent improvements to a PDB sequence culling server. Nucleic Acids Research 2005;33:W94–8.

Figure Captions

Figure 1. Average root-mean-square deviations (RMSD) in the C_α positions $\langle\Delta(i,j)\rangle$ (in Å) of all residues in the larger database of NMR structures without x-ray crystal structure pairs, plotted as a function of the number of restraints on each residue N_r. The average is taken over the multiple structures (~ 20) in each bundle.

Figure 2. (a) Probability distributions $P\ (\Delta_{core})$ of the root-mean-square deviations (RMSD) in the positions of the C_α atoms (in Å) for core residues in duplicate x-ray crystal structures (solid black line), in the NMR model ensemble for each structure (solid red line), and in paired x-ray crystal and NMR structures (dot-dashed blue line). We also plot the distribution for $\Delta = \sqrt{3B/8\pi^2}$ from the B-factor of core C_α atoms in the duplication x-ray crystal structures (dashed black line). The inset shows an example of one of the proteins in the paired x-ray crystal and NMR structure dataset, with the x-ray crystal structure on the left and the bundle of 20 NMR structures on the right (PDB codes 3K0M and 1OCA, respectively). The α -helices are colored purple, the β -sheets are yellow, and the loops are gray. (b) The fraction of core amino acids $F(\Delta\chi)$ with root-mean-square deviations of the side chain dihedral angles less than $\Delta\chi$ (in degrees) for the pairwise comparisons in (a). The inset is a schematic in two dimensions of the high-dimensional volume in configuration space that the C_α atoms in core residues in x-ray crystal structures and NMR ensembles sample. X-ray crystal structures sample a smaller region than NMR ensembles, but the distance *between* these regions of configuration space is larger than the fluctuations of both the x-ray crystal and NMR structures. The relative lengths of the arrows are drawn to scale, with $\langle\Delta_{core}\rangle \approx\ 0.1, 0.5, \text{and } 0.8$Å for the x-ray duplicates, NMR models, and pairs of x-ray crystal and NMR structures, respectively.

Figure 3. (a) Fraction of side chain conformations of core residues with predictions from the hard-sphere plus stereochemical constraint model that deviate from the experimentally observed values by less than $\Delta\chi$ (in degrees) in the dataset of x-ray crystal (solid black line) and NMR (solid red line) structure pairs, and the Dunbrack 1.0 dataset of 221 high resolution x-ray crystal structures (dashed black line) [9, 24]. (b) Fraction of core hydrophobic side chains, grouped by residue type, that can be predicted to within 30° of the corresponding experimental structure using the hard-sphere plus stereochemical constraint model for x-ray (black bars) and NMR structures (red bars). (c) Distribution of the overlap potential energy U_{RLJ}/ϵ, calculated using Eq.6 for core residues in the x-ray crystal (black line) and NMR structures (red line) in the paired dataset.

Figure 4. Distribution $P(\phi)$ of the packing fraction of core residues in the Dunbrack 1.0 dataset of high-resolution x-ray crystal structures (black dashed line), the dataset of high-resolution NMR structures for which there is not a corresponding x-ray crystal structure (red dashed line), and x-ray crystal structures (black solid line) and NMR structures (red solid) from the paired dataset.

Figure 5. Ensemble-averaged packing fraction $\langle\phi_J\rangle$ of jammed packings of amino-acid-shaped particles versus the dimensionless timescale τ for $N = 16$ particles. The colors represent simulations with different temperatures k_BT/ϵ, logarithmically spaced from 10^{-7} (blue) to 1 (red). The dashed black line at $\langle\phi_J\rangle = 0.55$ is the average packing fraction of core residues in x-ray crystal structures, and the dashed red line at $\langle\phi_J\rangle = 0.59$ is the average packing fraction of core residues in NMR structures.